DOCUMENT RESUME

ED 434 921 TM 030 172

AUTHOR Adair, James H.; Berkowitz, Nancy F.
TITLE Live Application Testing: Performance Assessment with Computer-Based Delivery.
PUB DATE 1999-04-00
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adults; *Certification; *Computer Assisted Testing; Computer Software; Job Skills; *Performance Based Assessment; Work Sample Tests
IDENTIFIERS *Lotus Notes

ABSTRACT

To measure workplace skills more realistically for certification purposes, two computer-delivered performance examinations, termed "Live Application" exams, were developed to test job-related competencies in a specific software product, Lotus Notes. As in the real world, success on examination tasks was determined by the examinee's final product, not interim keystrokes. A specially developed scoring program evaluated results of examination tasks. In the final examination format, examinees received a detailed score report immediately on completing the examination. The beta versions of the examinations were administered over a 3-month period to 171 people. A follow-up questionnaire was then sent to all of these beta examinees. Respondents reported feeling that the examination was fair (80% agreement) and a good test of their skills (90% agreement), and they also reported that taking the examination was a valuable experience (95% agreement). (Contains 5 tables and 18 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 434 921

Paper submitted to AERA 1999

Live Application Testing:

Performance Assessment with Computer-Based Delivery

James H. Adair and Nancy F. Berkowitz

Lotus Development Corporation

TM030172

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

James Adair

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Abstract

To more realistically measure workplace skills for certification purposes, two computer-delivered performance exams, termed "Live Application" exams, were developed to test job-related competencies in a specific software product, Lotus Notes. As in the real world, success on exam tasks was determined by the examinee's final product, not interim keystrokes. A specially developed scoring program evaluated results of exam tasks. In the final exam format, examinees receive a detailed score report immediately upon completing the exam. The beta exams were administered over a three month period to 171 people. A follow-up questionnaire (66% response rate) was then sent to all beta examinees. Respondents reported feeling the exam was fair (>80% agreement), was a good test of their skills (>90% agreement), and that taking the exam was a valuable experience (>95% agreement).

Those with the mandate to devise measures of people's skills are well aware of the pitfalls in such an undertaking. An efficient test of skill-related knowledge can be an inadequate measure of the performance actually desired of someone taking such a test. Schools and industry alike need assessment instruments that can be uniformly administered and objectively scored for those behaviors truly of interest. This paper describes an innovative methodology that combines performance assessment with computer-based assessment delivery. The paper includes a description of how the assessment tool was developed and how it works, and a report on its psychometric properties as well as examinee reactions in a pilot beta field test.

Over the last three years, a performance assessment technique called "Live Application" testing was developed to measure specific computer software competencies. The assessment technique requires examinees to perform a series of tasks on a computer, requiring actions to be taken rather than questions to be answered. The tasks measure application, analysis, and synthesis level skills, as well as procedural skills. The tasks were determined to be critical in accomplishing workplace tasks, based on job-task analyses.

Examinees perform exam tasks on stand-alone PCs connected to a computer network. Tests are downloaded from a remote site via the internet. After an examinee completes the examination tasks, computerized algorithms score the actions taken and evaluate the results for meaning. Results are sent back to the delivery vendor for quality audits and then to the test owner for exam maintenance and archiving. During the pilot exam period described in this paper, examinees received score notification after the exam owner had determined pass/no pass status. In the final exam form, the scoring program gives examinees instant feedback on their performance.

BEST COPY AVAILABLE

Review of the Literature

Performance assessment has been described as a means to demonstrate understanding and skills in applied, procedural, or open-ended settings (Baker, 1993; Airasian, 1994). Performance assessments are further distinguished by their ability to show what a person can do in a real situation (Fitzpatrick and Morrison, 1971; Wiggins, 1992), in contrast to multiple choice examinations whose questions are often deemed inadequate to assess what test-takers can do and what they know (Gibbs & Peck, 1995). Performance assessments have been used in a number of ways including diagnosis and remediation, formative evaluation, and certification (Baker, 1993). They are also seen as instruments of standards based education (Messick, 1989).

Computer-based performance examinations have been used to measure behaviors in a variety of content areas including components of understanding, team problem solving, medical student clinical performance, science abilities, and competence in areas of Information Technology, among others (Baker & O'Neil, 1995; Fitzgerald et al., 1994; Kumar & Helgeson, 1995; Malone & Barry, 1996). They have also been used in situated learning with science and with thinking processes (Young, 1995). Issues significant in many performance examinations such as inter-rater reliability (Klein et al., 1998) and subjectivity in scoring (Braun et al., 1990) may be less significant or even sidestepped by the computer-based testing and scoring. Certain shortcomings of computer-based evaluation have been identified, such as a computer's difficulty evaluating complex responses and gender equity issues (Gibbs & Peck, 1995; Cheek & Agruso, 1995). Because of the unique role of the computer in Information Technologies, computer-based delivery of performance certification exams allows exam tasks to be virtually identical to real-world job tasks, thus potentially ameliorating the above concerns.

Evidence supporting validity of certification or licensure exams hinges on the appropriateness of exam content, and this content is defined primarily through job analyses (AERA, APA, & NCME, 1985). Once this certification and professional licensure content is carefully defined, candidates are often tested with computerized multiple choice examinations. The advantages of multiple choice examinations include the potential for worldwide test delivery and immediate score reporting. Such advantages of multiple choice tests also need to be captured by computer-based performance assessment.

Certification examinations in the Information Technologies field are oriented toward specific Information Technologies products, so testing candidates' actual skills using these products is a worthwhile goal. Computer-based performance certification examinations are not new to this field. Some exams use simulations to measure specific Information Technology competencies (Foster, 1996). Performance exams where examinees work within the product have been developed by other Information Technology vendors (Malone & Barry, 1996; G. Yeung, former manager of Professional Certification, Sybase, Inc., personal communication, October 1997). However, these in-product exams suffer from the lack of an immediate scoring feedback mechanism (Yeung) or from a limited geographic market penetration (Malone & Barry).

An ideal computer-based performance assessment for certification in the Information Technologies field, and perhaps other fields as well, would be amenable to worldwide test

BEST COPY AVAILABLE

delivery and immediate score reporting. It would capture a representative, real-world set of tasks enabling the certifying organization, relevant employers, and examinees to conclude that passing the examination was a worthwhile and meaningful standard.

Methodology

The Live Application computer-based assessment tool was developed for use in certifying individuals in the Information Technology field. Development took place over three years. The development consisted of three distinct aspects: the tasks to be performed, the scoring mechanism, and the exam delivery mechanism.

A. Task Definition

Formal Job Task Analysis (JTA) anchored this assessment tool by identifying essential competencies. Specific tasks were drawn from the JTA-identified competencies, and then these tasks were converted into an operational performance definition. Performance in these exams was defined as the final outcome versus detailed interim steps, as it is the final outcome that is germane to the workplace skills these exams were designed to measure.

Two Live Application exams were built in 1996 by the psychometric services team of Lotus Development Corporations' certification program. One exam assesses skills needed for the development of applications within the Lotus Notes environment. The other assesses the ability to administer a Lotus Notes system. The Live Application exams require a person to perform a series of tasks within an integrated performance scenario. These are divided into a number of sub-tasks. Examinees use the actual product in which they seek certification (Lotus Notes) within a computer environment typically used for the assessed skill set.

Each of the two exams consists of a list of tasks to be performed and one or more databases in which to perform the tasks. Those performance tasks reflect real-world job situations because databases are the environment for the product competencies being tested. An examinee is presented with a task to perform and then goes to the appropriate database to perform that task.

The Lotus Notes Applications Development exam requires the examinee to build input screens (Forms), display screens (Views), and small programs (Agents) which change data. The following is an example similar to the test scenario. Examinees create database elements to store and display product and pricing information for a sporting goods company. Examinees create product order forms that include design elements ranging from product suggested retail price to a specified input screen color. They create elements, tailored to specific user input, to automatically move a database user between input screens or from the database itself. Examinees then create efficient display screens for grouping input fields. They set up appropriate user access to the database, and they create small programs that change data as needed. Examinees also troubleshoot and fix preexisting errors in the databases.

The Lotus Notes Systems Administration exam requires the examinee to complete Forms which define how the computer system is to work, and to perform tasks which affect the

BEST COPY AVAILABLE

system's performance. To give an example similar to the exam, examinees expand and build upon an existing system with four servers. Examinees modify these servers as well as create a new certification scheme, new servers, and new users. The examinee also troubleshoot and fix preexisting errors in Forms and system performance.

Exam outcome-measure data points consist of keywords and formulae supplied by examinee performance on specified exam tasks. Outcome measures of examinee tasks are direct except in four specific incidences where the task did not allow for direct measurement. In those cases, the examinee selects from a list of up to 20 different alternatives, which simulate how that task would be performed in a real-world situation using the actual product.

Each of these exams was tested to determine task appropriateness and wording as well as test length. Examinees were allowed three hours to complete the beta exam. After that time or after completing all tasks, the examinee would terminate the exam. Beta exams were hand scored to determine answer variability and resulting scoring rules. In the final exam format, examinees immediately receive a detailed score report.

B. Scoring Mechanism

Once the exam performance was defined, the scoring mechanism and algorithms were created. The scoring mechanism was developed in three stages: the desired performance outcomes were defined; the rules for measuring that outcome were established; and a program was written to determine whether the defined outcome had been reached.

The desired performance outcomes. The desired performance outcomes were defined so that scores from zero to 100 percent were possible. First, the starting performance environment was established as a zero percent baseline, so that examinees who performed no tasks during the examination period received a zero score. Second, exam tasks were performed one at a time. The result was a 100 percent solution, so that if examinees performed each task correctly during the examination period they received a score of 100.

The scoring rules. Two decisions were made regarding the scope of measurement. First, only those tasks completed correctly are measured, and examinee performance unrelated to the desired outcome are ignored. Second, performances are measured at the sub-task level rather than at the task level, so that it is possible to give partial credit for tasks attempted.

Each task is evaluated to see how the correct completion of that task could be scored. The scoring program looks into the database where the tasks are performed. It pulls out the performance and compares it to the correct performance using either a direct or indirect rule. The direct scoring rule compares a task performance to the predefined correct answer or answers. The indirect scoring rule takes the results of a task performance and runs it against background data to determine whether the performance delivers an acceptable solution. The scoring program evaluates each point on a success/nonsuccess basis. The total number of performance points correct is compared to the total number of performance points possible. The total score as well as the performance points are passed back to the delivery system so that, after the beta exam period, examinees receive immediate feedback through a score report.

BEST COPY AVAILABLE

The scoring program: The scoring program, written by a third party development group, took the shape of an Applications Program Interface (API) program. An API program can score examinee performance because of the object structure of the software program in which the examinee works, Lotus Notes. Each aspect of a Notes database is an object with a unique identifier. This structure allows for identification of an object and the object attributes. Since Live Application exam tasks involve the creation of objects along with the setting of appropriate attributes, the API scoring program can search through the results of examinee performance, examining one or more database objects modified as a result of that performance. The scoring algorithm judges the success and/or failure of the performance in one of two ways. The first way is to look for keywords in the object attributes that should appear in specific locations. The second way is to take keyword sequences from the object attributes and run them against background data to see if the prescribed results are obtained. The results are recorded in a success (1) or non success (0) format.

C. Delivery Mechanism

The delivery mechanism presents the examinee with the tasks to be performed and the databases in which to perform the tasks. The examinee carries out these tasks within a live session of the actual software. The results of this performance were, in the beta exam period, sent to the exam owner for scoring. The pass/no pass score was determined by subject matter experts, using a modified Angoff technique (Livingston & Zieky, 1982), a technique found to be utilized frequently in setting performance standards (Plake, 1998).

Although in the initial exam pilot period examinees received no on-screen score report, in the current exam form, examinees receive a score report immediately upon finishing. That score report indicates the number and percentage of correctly answered performance points, accompanied by an overall pass/no pass rating. The report includes detailed exam section reports that describe the percentage of scored performance points related to the various exam tasks. Finally, the score report presents a list of competencies related to those tasks where incorrect responses were recorded.

The beta exams were delivered in a controlled number of worldwide testing sites at locations in North America, Europe, and Asia, using an international test delivery vendor's network of exam sites. Exams were proctored.

The number of exam sites was determined by the availability of computers able to handle the Live Application performance assessments. The ability to handle these exams was defined by a set of minimum hardware and software requirements. The programming and administration environment (Lotus Notes) was installed and running on those machines. The computing environment was as close to the actual development or administration environment as possible. Thus appropriate Help files were available to look up information. However, "Joe" down the hall was not available to answer questions.

BEST COPY AVAILABLE

A follow-up questionnaire was mailed to all beta test-takers. The survey asked questions regarding the fairness and appropriateness of the exam, and it gathered background information on test-takers' exam related-experience and exam preparation methods.

Results

The Live Application exams were administered worldwide to 171 people over a three month period. Exam results and questionnaire results were analyzed using basic statistics and reliability procedures. Questionnaire and test data were combined using stepwise regression analysis to determine exam result predictors.

Test analysis included traditional reliability analyses. One could argue that traditional reliability analysis is not ideal for Live Application exams, since traditional analyses assume independence of items. Although major Live Application exam tasks were generally independent from each other, the sub-tasks within specific tasks were interdependent: Examinees could not hope to complete a sub-task if they had not successfully completed the first step. With that caveat and with a goal of evaluating the internal consistency of the exam to the degree possible with standard methods, traditional results are reported below.

Statistical analysis indicated high whole test internal consistency reliabilities, Chronbach alpha, of greater than .90. In addition, subtest reliabilities greater than .80 were found for subtest scores consisting of as few as three items (insert Tables 1 and 2 here). The very high overall test internal consistency values are noteworthy despite inflation by certain sub-task interdependencies. This high degree of internal consistency occurred despite indications that the beta testing environment was not ideal. Several examinees commented on the follow-up questionnaire that test directions needed to be improved.

The follow-up questionnaire sent to the Live Application examinees (66% return rate), indicated the exam was both fair (80% agreement) and a better measure of their skills than multiple choice exams (90% agreement) (insert Table 3 here). It must be noted that, due to the beta nature of the exam, many completed their follow-up questionnaires providing positive feedback before knowing whether they passed the exam. Thus, positive feedback was not necessarily based on receiving high exam results. Stepwise regression analysis indicated that over 40% of the variance could be accounted for by experience, preparation, and/or motivation (insert Tables 4 and 5 here).

Perhaps the most important result to report is the successful construction of two computer delivered performance exams, exams that examinees felt were both fair and a good test of their skills, in which examinee performance was objectively scored. Examinee performance consisted of correct object manipulation in a computer-based object store, in this instance Lotus Notes. The attributes and content in the object store contain the results of a person's performance. Retrieving that information and comparing it to scoring rules allows creation of objective performance scores.

BEST COPY AVAILABLE

IV. Conclusions

During a three month period in 1996, live application performance assessments were given to 171 applicants for certification in either applications development or systems administration. Both assessments were found to be reliable (>.90 coefficient alpha) measures of candidate performance. Furthermore, candidates reported that they believed the assessment to be fair (>80% agreement) and the best way to test their skills (90% agreement). A stepwise regression analysis revealed that over 40% of the variance on each exam could be accounted for by exam task-related experience, preparation and/or motivation. The exams prevented guessing as they required completion of in-product real-world job tasks. These exam tasks were based on Job Task Analyses, a primary means for defining the content domain. Validity evidence for these certification exams stemmed from careful selection of exam content. Due to the nature of the exam tasks performed, the tests were also perceived to be highly valid.

Discussion and Implications

The above results were based on worldwide distribution of the exam. Those results, however, were based on the relatively small number of people who took the Live Application exam in the beta period. Although small numbers are characteristic of many Information Technologies certification exams, the small sample is an obvious limitation to the study. Further discussion of the study focuses on the Live Application exams, themselves.

The large number of data points gathered using this technique is quite large compared to the number of data points that can be gathered using a more conventional "multiple choice" type format. In the Information Technology certification industry, a rule of thumb is to allow one and a half minutes to answer each multiple-choice type item. In addition, we try to restrict exams to a 60 minutes time limit, thus resulting in an exam of approximately 40 knowledge/comprehension level items giving us 40 data points. An alternative would be to use a multiple-multiple choice format allowing for 160 knowledge/comprehension level data points. However this format results in customer dissatisfaction with the certification experience. On the other hand, we can gather approximately 40 performance/procedural data points within a 15 minute period using the Live Application technique. An exam which lasts an hour can produce 150-200 performance/procedural data points.

The ability to gather many performance data points within a short time span allows us to create testlets which can be combined with other testlets to produce "situational" exams. In addition, the large number of data points derived by using this technique can also be used to diagnostically assess and, in turn, prescribe remediation.

There are three major limitations of the Live Application exams. The integrated performance scenario format introduces potential areas of bias. The first is a scenario bias introduced by choosing any specific scenario for an exam. The scenario bias is ameliorated by having the exam scenario match the intent of the job being measured. In this instance each exam used a scenario similar to the one used in a major accompanying learning resource.

BEST COPY AVAILABLE

The use of a carefully constructed performance exam scenario lent itself to establishing performance standards in a holistic manner. We were able to ask, "if a person does the prescribed performance, would this be accepted as evidence that the person can do the requisite job?" This holistic technique and JTA based content supported the validity of the Live Application exams. Multiple choice exams in Information Technologies are more limited in approaches to setting performance standards.

The second and third limitations involve bias arising from potential areas of dependency. Both item scoring dependency and task dependency are potential problems with the Live Application exams. Item scoring dependency occurs when the scoring algorithm contains the same dependency for two or more items. In the Live Application exams, the scoring program includes a common path to find the performance for some data points. If the start of the path is missed, all data points along that path would also be missed. Although we adjusted the scenario to avoid such occurrences in the current exams, a common path in a live application performance exam could lead to spuriously high reliabilities.

Task inter-item dependency is an even greater threat to reliability. This statistical problem reflects the real world, however, as a programmer builds a highly integrated piece of software where tasks are interrelated. For example, if a developer spells something incorrectly early on in a program and the program later does not work, it becomes very difficult to determine exactly what went wrong and every aspect of the program that was subsequently affected. Thus task interdependency is operative in the jobs for which certification through these exams is sought. However, inter-item dependencies can introduce a bias affecting reported reliability.

A strength of Live Application performance exams is that they take place within a real-world context. That context can be adapted to measure and diagnostically assess any skills which can be demonstrated within a computer setting. Their greatest strength is in handling procedural skills as well as those skills that involve application, analysis, and synthesis.

If one has spent time defending high-stakes multiple choice exams, supporting Live Application exams is a welcome change of pace. With multiple choice exams, there are often complaints about a particular question's clarity, or accuracy or appropriateness for an examinees' specific job. One hears far fewer complaints of Live Application exams. Contrasted with trying to explain why a multiple choice item was important to a job-related exam, Live Application examinees often comment, "this is just like doing my job." While an examinee not passing a multiple choice exam may be frustrated about how to prepare to retake the exam, examinees know to prepare to retake a Live Application exam through more job experience.

Setting a passing score on a multiple choice exam seems, even in the best of circumstances, arbitrary in comparison to determining minimal competence on Live Application performance examinations. The examinee also obtains far superior

BEST COPY AVAILABLE

post-exam feedback, as detailed score reports can reflect weaknesses on accomplishing job-related tasks versus failure to answer a group of multiple choice questions.

V. Educational Importance

This performance assessment technique has implications beyond certification in the Information Technology industry. Conceptually, the Live Application technique can be used to assess the results of a standards-based education. It can be used to diagnostically assess individual strengths and weaknesses. The Live Application technique can be used as a means to assess skills and competencies on both a formative and summative basis and thus provide a means for computerized adaptive learning.

The same Live Application examination environment used to assess performance in Lotus Notes could be used to assess other skills, assuming the assessment took place within a specific object language or product for which an API program was developed. Results of answers to mathematics problems could be captured by a modification of the Live Application scoring algorithm. This same performance measurement technique could be used to measure recognition level grammar, punctuation, and spelling skills such as are found on standardized student achievement and teacher certification exams. It could also be used to assess programming skills in various programming languages. The success of the methodology developed for the Live Application exams suggests the attainability of objective computer-based measures of performance in various content areas.

BEST COPY AVAILABLE

Table 1

Reliability of Application Development Subtests

Logical Tests and Subtests for Application Development Live Application Assessment	Number of Examinees	Number of Items	Reliability
Application Development I	**96**	**122**	**0.98**
Views		52	0.97
Columns Properties		32	0.97
View Properties		13	0.87
Column formula		7	0.85
Fields		44	0.97
Data & Field Type		15	0.92
Field Name		13	0.92
Keyword Fields		7	0.90
Inheritance, computed & editable fields		9	0.86

BEST COPY AVAILABLE

Table 2

Reliability of System Administration Subtests

Logical Tests and Subtests for System Administration Live Application Assessment	Number of Examinees	Number of Items	Reliability
System Administration I	**75**	**100**	**0.98**
Install & Set up		47	0.98
Install & Set up task A		37	0.95
Install & Set up task A1		14	0.95
Install & Set up task A2		8	0.97
Install & Set up task A3		7	0.86
Install & Set up task A4		6	0.99
Install & Set up task B		10	0.82
Install & Set up task B1		6	0.83
Install & Set up task B2		4	0.94
Install & Set up Server Documents		28	0.97
Install & Set up Server task A		12	0.97
Install & Set up Server task B		5	0.72
Install & Set up Server task C		4	0.94
Install & Set up Server task D		5	0.92
Maintenance & Operations		16	0.88
Maintenance & Operations A		11	0.85
Maintenance & Operations B		5	0.87
Troubleshooting		4	0.72
Systems Security		5	0.74

BEST COPY AVAILABLE

Table 3

Responses to Live Application Exam Follow-up Questionnaire

	Application Development		System Administration	
Item	**Percent Agree**	**Percent Disagree**	**Percent Agree**	**Percent Disagree**
Exam tested skills	95	3	94	3
Exam was fair	81	11	83	11
Exam was a valuable experience	96	0	95	5
Live Application is the best way to test skills	90	6	90	8
Live Application increases confidence in the certification process	80	9	82	8
Multiple choice exams test skills better	17	73	11	76

BEST COPY AVAILABLE

Table 4

Regression Results for Application Development Percent Score

Predictor	**Beta**	**Significance**
Time using Lotus exam guides	-0.822	0.000
Reason for taking exam - personal satisfaction	-0.276	0.011
Used Lotus education exam guides	0.503	0.018
Reason for taking exam - new career	-0.252	0.023
Adjusted R square = 0.42		
F = 12.2		
P < 0.01		
degrees freedom = 4,57		
n = 61		

BEST COPY AVAILABLE

Table 5

Regression Results for System Administration Percent Score

Predictor Variables	**Beta**	**Significance**
Reason for taking exam - new career	-0.298	0.006
Reason for taking exam - personal satisfaction	-0.251	0.017
Programming background in an object language	-0.305	0.002
Reason for taking exam - increased salary	-0.212	0.026
Hours per week on the job experience	0.213	0.030
Adjusted R square = 0.54		
F = 12.0		
P < 0.01		
degrees of freedom = 6,51		
n = 57		

BEST COPY AVAILABLE

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Airasian, P. (1994). Classroom assessment. New York: McGraw Hill.

Baker, E. L., O'Neil, H. F., (1995). Computer technology futures for the improvement of assessment. Journal of Science Education and Technology, 4(1), 37-45.

Baker, E. L., O'Neil, H. F., Jr. & Linn, R. L., (1993). Policy and validity prospects for performance-based assessment. American Psychologist, 42, 1210-1218.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27(2), 93-107.

Cheek, D. W., Agruso, S. (1995). Gender and equity issues in computer-based science assessment. Journal of Science Education and Technology, 4(1), 75-79.

Fitzgerald, J. T., Wolf, F. M., Davis, W. K., Barclay, M. L., Bozynaski, M. E., Chamberlain, K. R., Clyman, S. G., Shope, T. C., Woolliscroft, J. O., Zelenock, G. B., A Preliminary study of the impact of case specificity on computer-based assessment of medical student clinical performance. Evaluation and the Health Professions, 17(3), 307-321.

Fitzpatrick, R. & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.) Educational Measurement (pp. 237-270). Washington, D.C.: American Council on Education.

Foster, D. (1996, October). New technologies for performance-based testing. Paper Presented at the Certification 96 Conference: Blueprint for Leadership, Houston, Texas.

Gibbs, W. J., Peck, K. L. (1995). An approach to designing computer-based evaluation of student constructed responses: Effects on achievement and instructional time. Journal of Computing in Higher Education, 6(2), 99-119.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Scoring performance assessments. Applied Measurement in Education, 11(2), 121-137.

Kumar, D. D., & Helgeson, S. L. (1995). Trends in computer applications in science assessment. Journal of Science Education and Technology, 4(1), 29-36.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Educational Testing Service.

BEST COPY AVAILABLE

Malone, E., & Berry, T. (1996, October). Computerized performance testing: Forward into the past. Paper Presented at the Certification 96 Conference: Blueprint for Leadership, Houston, Texas.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.). (pp. 13-103). New York: Macmillan.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. Applied Measurement in Education, 11(1), 65-79.

Wiggins, G. (1992). Creating tests worth taking. Educational Leadership, 49(8), 26-33.

Young, M. (1995). Assessment of situated learning using computer environments. Journal of Science Education and Technology, 4(1), 89-96.

BEST COPY AVAILABLE